Numerical Optimisation: Constraint Optimisation

Marta M. Betcke
m.betcke@ucl.ac.uk,

Kiko Rullan
f.rullan@cs.ucl.ac.uk

Department of Computer Science,
Centre for Medical Image Computing,
Centre for Inverse Problems
University College London

Lecture 12

Constraint optimisation problem

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad \begin{cases} c_i(x) = 0, & i \in \mathcal{E} \\ c_i(x) \geq 0, & i \in \mathcal{I} \end{cases} \qquad \text{(COP)}$$

- $f : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}$: objective function, assume smooth
- $c_i : \mathbb{R}^n \to \mathbb{R}$: constraint function, assume smooth
  $i \in \mathcal{E}$ equality constraints,
  $i \in \mathcal{I}$ inequality constraints.

- $x \in \mathbb{R}^n$: optimisation variable

Feasible set Ω is a set of all points satisfying the constraints

$$\Omega = \{x \in \mathcal{D} : \; c_i(x) = 0, \; i \in \mathcal{E}; \; c_i \geq 0, \; i \in \mathcal{I}\}.$$

Optimal value: $x^\star = \inf_{x \in \Omega} f(x)$

- $x^\star = \infty$ if (COP) is infeasible i.e. $\Omega = \emptyset$
- $x^\star = -\infty$ if (COP) is unbounded below

Smooth constraints can describe regions with *kinks*.

Example: 1-norm:

$$\|x\|_1 = |x_1| + |x_2| \le 1$$

can be described as

$$x_1 + x_2 \le 1, \quad x_1 - x_2 \le 1, \quad -x_1 + x_2 \le 1, \quad -x_1 - x_2 \le 1.$$

Example: pointwise max

$$\min f(x) = \max(x^2, x)$$

can be reformulated as

$$\min t, \quad \text{s.t.} \quad t \ge x, \quad t \ge x^2.$$

Types of minimisers of constraint problems

A point $x^\star \in \Omega$ is a **global minimiser** if

$$f(x^\star) \leq f(x), \ \forall x \in \Omega.$$

A point $x^\star \in \Omega$ is a **local minimiser** if

$$\exists \mathcal{N}(x^\star) : f(x^\star) \leq f(x), \ \forall x \in \mathcal{N}(x^\star) \cap \Omega.$$

A point $x^\star \in \Omega$ is a **strict (or strong) local minimiser** if

$$\exists \mathcal{N}(x^\star) : f(x^\star) < f(x), \ \forall x \in \mathcal{N}(x^\star) \cap \Omega, \ x \neq x^\star.$$

A point $x^\star \in \Omega$ is an **isolated local minimiser** if

$$\exists \mathcal{N}(x^\star) : x^\star \text{ is the only local minimiser in } \mathcal{N}(x^\star) \cap \Omega.$$

Active set

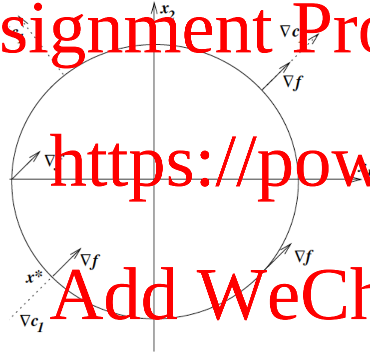
Feasibility problem: Find x such that all constraints are satisfied at x.

Active set $\mathcal{A}(x)$ at any feasible $x \in \Omega$ consists of the equality constraint indices set $\mathcal{E}$ and the inequality constraints $i \in \mathcal{I}$ for which $c_i(x) = 0$

$$\mathcal{A}(x) = \mathcal{E} \cup \{i \in \mathcal{I} : \; c_i(x) = 0\}.$$

At a feasible point $x \in \Omega$, the inequality constraint $i \in \mathcal{I}$ is said to be *active* if $c_i(x) = 0$ and *inactive* if the strict inequality holds $c_i(x) > 0$.

$$\min x_1 + x_2 \quad \text{s.t.} \quad x_1^2 + x_2^2 - 2 = 0.$$



Feasibility: (Taylor expansion of c_1)

$$0 = c_1(x+s) \approx \underbrace{c_1(x)}_{=0} + \nabla c_1(x)^{\mathrm{T}} s$$

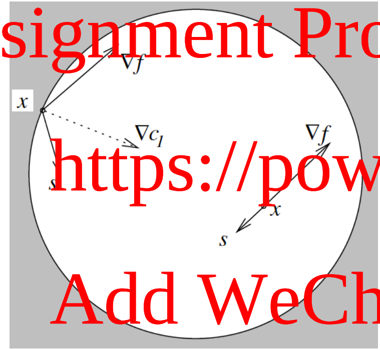
Decrease direction: (Taylor expansion of f)

$$0 > f(x+s) - f(x) \approx \nabla f(x)^{\mathrm{T}} s$$

The only situation that such s does not exist is if for some scalar λ_1

$$\nabla f(x) = \lambda_1 \nabla c_1(x).$$

$$\min x_1 + x_2 \quad \text{s.t.} \quad 2 - x_1^2 - x_2^2 \geq 0.$$



Feasibility: (Taylor expansion of c_1)

$$0 \leq c_1(x+s) \approx c_1(x) + \nabla c_1(x)^{\mathrm{T}} s$$

Decrease direction: (Taylor expansion of f)

$$0 > f(x+s) - f(x) \approx \nabla f(x)^{\mathrm{T}} s$$

Case: x inside the circle, i.e. $c_1(x) > 0$

$$s = -\alpha \nabla f(x)$$

Single inequality constraint

$$\min x_1 + x_2 \quad \text{s.t.} \quad 2 - x_1^2 - x_2^2 \geq 0.$$

Feasibility: (Taylor expansion of c_1)

$$0 \leq c_1(x+s) \approx \underbrace{c_1(x)}_{=0} + \nabla c_1(x)^{\mathrm{T}} s$$

Decrease direction: (Taylor expansion of f)

$$0 > f(x+s) - f(x) \approx \nabla f(x)^{\mathrm{T}} s$$

Case: x on the boundary of the circle, i.e. $c_1(x) = 0$

$$\nabla f(x)^{\mathrm{T}} s < 0, \quad \nabla c_1(x)^{\mathrm{T}} s \geq 0$$

Empty only if $\nabla f(x) = \lambda_1 \nabla c_1(x)$ for some $\lambda_1 \geq 0$.

Given the point x in the active set $\mathcal{A}(x)$, the **linear independent constraint qualification (LICQ)** holds if the set of of active constraint gradients $\{\nabla c_i(x),\ i \in \mathcal{A}(x)\}$ is linearly independent.

Note that for LICQ to be satisfied, none the the active constraint gradients can be 0.

Example: LICQ is not satisfied if we define the equality constraint $c_1(x_2^2 + x_2^2 - 2)^2 = 0$ (same feasibility region, different constraint)

There are other constraint qualifications e.g. Slater's conditions for convex problems.

Lagrangian function

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x).$$

Let $x^\star$ be a local solution of (COP) and f and c_i be continuously differentiable and LICQ hold at $x^\star$. Then there exists a **Lagrange multiplier** $\lambda^\star$ with components $\lambda_i^\star$, $i \in \mathcal{E} \cup \mathcal{I}$ such that the following **Karush-Kuhn-Tucker conditions** are satisfied at $(x^\star, \lambda^\star)$

$$\nabla_x \mathcal{L}(x^\star, \lambda^\star) = 0 \tag{1a}$$

$$c_i(x^\star) = 0, \quad \forall i \in \mathcal{E}, \tag{1b}$$

$$c_i(x^\star) \geq 0, \quad \forall i \in \mathcal{I}, \tag{1c}$$

$$\lambda^\star \geq 0, \quad \forall i \in \mathcal{I}, \tag{1d}$$

$$\lambda_i^\star c_i(x^\star) = 0, \quad \forall i \in \mathcal{E} \cup \mathcal{I}. \tag{1e}$$

Strong complementarity condition

The *complementarity condition* (2)(e) can be made stronger.

Given $x^\star$ a local solution of (COP) and a vector $\lambda^\star$ satisfying the KKT conditions (2), the **strict complementarity condition** holds if exactly one of $\lambda_i^\star$ and $c_i(x^\star)$ is zero for each $i \in \mathcal{I}$. In other words, $\lambda_i^\star > 0$ for each $i \in \mathcal{I} \cap \mathcal{A}(x^\star)$.

Strict complementarity makes it easier for the algorithms to identify the active set and converge quickly to the solution.

For a given solution $x^\star$ of (COP), there may be many vectors $\lambda^\star$ which satisfy the KKT condition (2). However, if LICQ holds, the optimal $\lambda^\star$ is unique.

For convenience we change (and refine) our notation for the constraint optimisation problem. The following slides are based on Boyd (Convex Optimization I).

Let p^* be the optimal value of the **primal problem**

$$\begin{aligned} \min_{x \in \mathcal{D} \subset \mathbb{R}^n} \quad & f(x) && \text{(COP:P)} \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{aligned}$$

The **Lagrangian** $\mathcal{L} : \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$

$$\mathcal{L}(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

- λ_i are Lagrange multipliers associated with $f_i(x) \leq 0$
- ν_i are Lagrange multipliers associated with $h_i(x) = 0$

Lagrange dual function: $g: \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda, \nu) \quad \text{(LD)}$$

$$= \inf_{x \in \mathcal{D}} \left(f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right).$$

g: is concave, can be $-\infty$ for some λ, ν.

Lower bound property: If $\lambda \geq 0$, then $g(\lambda, \nu) \leq p^\star$.
Proof: For any feasible $\tilde{x}$ and $\lambda \geq 0$ we have

$$f(\tilde{x}) \geq \mathcal{L}(\tilde{x}, \lambda, \nu) \geq \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda, \nu) = g(\lambda, \nu).$$

Minimising over all feasible $\tilde{x}$ gives $p^\star \geq g(\lambda, \nu)$.

Convex problem

Convex problem in standard form

$$\begin{aligned} \min_{x \in \mathcal{D} \subset \mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \dots, m, \\ & h_i(x) = 0, \quad i = 1, \dots, p \end{aligned}$$

- f is convex and $\mathcal{D}$ is convex
- f_i are convex
- h_i are affine i.e. $a_i^T x = b_i$

Feasibility set Ω of a convex problem is a convex set.

$$\min_{x\in\mathbb{R}^n} \quad x^{\mathrm{T}}x$$
$$\text{subject to} \quad Ax = b$$

- Lagrangian: $\mathcal{L}(x,\nu) = x^{\mathrm{T}}x + \nu^{\mathrm{T}}(Ax - b)$
- Dual function:
  $g(\nu) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x,\nu) = \inf_{x\in\mathbb{R}^n} \left(x^{\mathrm{T}}x + \nu^{\mathrm{T}}(Ax - b)\right)$

- $\mathcal{L}(x,\nu)$ is strictly convex in x, to minimise over x set gradient equal zero:

$$\nabla_x \mathcal{L}(x,\nu) = 2x + A^{\mathrm{T}}\nu = 0 \quad \Rightarrow \quad x_{\min} = -1/2A^{\mathrm{T}}\nu$$

- Plug $x_{\min}$ into g

$$g(\nu) = \mathcal{L}(x_{\min},\nu) = -\frac{1}{4}\nu A^{\mathrm{T}}A\nu - b^{\mathrm{T}}\nu.$$

g is a concave function of ν.

Lower bound property: $p^\star \geq -1/4\nu A^{\mathrm{T}}A\nu - b^{\mathrm{T}}\nu$ for all ν.

Example: standard form LP

$$\min_{x\in\mathbb{R}^n} \quad c^{\mathrm{T}}x$$
$$\text{subject to} \quad Ax = b, \quad x \geq 0$$

- Lagrangian:
  $\mathcal{L}(x,\nu) = c^{\mathrm{T}}x + \nu^{\mathrm{T}}(Ax-b) - \lambda^{\mathrm{T}}x = -b^{\mathrm{T}}\nu + (c + A^{\mathrm{T}}\nu - \lambda)^{\mathrm{T}}x$

- Dual function:
  $$g(\lambda,\nu) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x,\lambda,\nu) = \inf_{x\in\mathbb{R}^n} \left(-b^{\mathrm{T}}\nu + (c + A^{\mathrm{T}}\nu - \lambda)^{\mathrm{T}}x\right)$$

- $\mathcal{L}(x,\nu)$ is affine in x, hence
  $$g(\lambda,\nu) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x,\nu) = \begin{cases} -b^{\mathrm{T}}\nu, & A^{\mathrm{T}}\nu - \lambda + c = 0 \\ -\infty, & \text{otherwise} \end{cases}$$
  g is linear on affine domain $\{(\lambda,\nu) : A^{\mathrm{T}}\nu - \lambda + c = 0\}$, hence concave.

Lower bound property: $p^\star \geq -b^{\mathrm{T}}\nu$ if $A^{\mathrm{T}}\nu + c \geq 0$.

$$\min_{x\in\mathbb{R}^n} \quad \|x\|$$
$$\text{subject to} \quad Ax = b$$

- Lagrangian:
$$\mathcal{L}(x,\nu) = \|x\| + \nu^{\mathrm{T}}(b - Ax) = \|x\| + b^{\mathrm{T}}\nu - \nu^{\mathrm{T}}Ax$$

- $\inf_{x\in\mathbb{R}^n}(\|x\| - y^{\mathrm{T}}x) = 0$ if $\|y\|_\star \le 1$, $-\infty$ otherwise, where $\|v\|_\star = \sup_{\|u\|\le 1} u^{\mathrm{T}}v$ is dual norm of $\|\cdot\|$.
If $\|y\|_\star \le 1$, then $\|x\| - y^{\mathrm{T}}x \ge 0, \forall x$, with equality if $x = 0$.
If $\|y\|_\star > 1$, choose $x = tu$, $u: \ \|u\| \le 1$, $u^{\mathrm{T}}y = \|y\|_\star > 1$
$$\|x\| - y^{\mathrm{T}}x = t(\|u\| - \|y\|_\star) \to -\infty \quad \text{as } t \to \infty$$

- Dual function:
$$g(\nu) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x,\nu) = \begin{cases} b^{\mathrm{T}}\nu, & \|A^{\mathrm{T}}\nu\|_\star \le 1 \\ -\infty, & \text{otherwise} \end{cases}$$

Lower bound property: $p^\star \ge b^{\mathrm{T}}\nu$ if $\|A^{\mathrm{T}}\nu\|_\star \le 1$.

Conjugate function

The **conjugate** of function f is

$$f^\star(y) = \sup_{x \in \mathcal{D}} (y^{\mathrm{T}} x - f(x))$$

The conjugate $f^\star$ is convex (even if f is not)

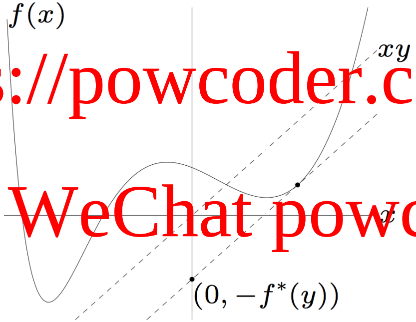


Figure: Boyd, Convex Optimization I

Lagrange dual and conjugate function

$$\min_{x \in \mathbb{R}^n} \quad f(x)$$
$$\text{subject to} \quad Ax \leq b, \quad Cx = d$$

- Lagrangian:

$$\begin{aligned}\mathcal{L}(x, \lambda, \nu) &= f(x) + \lambda^{\mathrm{T}}(Ax - b) + \nu^{\mathrm{T}}(Cx - d) \\ &= f(x) + (A^{\mathrm{T}}\lambda + C^{\mathrm{T}}\nu)^{\mathrm{T}}x - b^{\mathrm{T}}\lambda - d^{\mathrm{T}}\nu.\end{aligned}$$

- Dual function:

$$\begin{aligned}g(\lambda, \nu) &= \inf_{x \in \mathcal{D}} \left(f(x) + (A^{\mathrm{T}}\lambda + C^{\mathrm{T}}\nu)^{\mathrm{T}}x - b^{\mathrm{T}}\lambda - d^{\mathrm{T}}\nu\right) \\ &= -\sup_{x \in \mathcal{D}} \left(-f(x) - (A^{\mathrm{T}}\lambda + C^{\mathrm{T}}\nu)^{\mathrm{T}}x - b^{\mathrm{T}}\lambda - d^{\mathrm{T}}\nu\right) \\ &= -f^{\star}(-A^{\mathrm{T}}\lambda - C^{\mathrm{T}}\nu) - b^{\mathrm{T}}\lambda - d^{\mathrm{T}}\nu\end{aligned}$$

Lagrange dual problem

$$\begin{aligned} \max \quad & g(\lambda, \nu) \\ \text{subject to} \quad & \lambda \geq 0 \end{aligned}$$

- finds the best lower bound on p^* obtained from Lagrange dual function
- is a convex optimization problem, we denote its optimal value with d^*
- λ, ν are dual feasible if $\lambda \geq 0$, $(\lambda, \nu) \in \operatorname{dom} g$
- often simplified by making implicit constraint $(\lambda, \nu) \in \operatorname{dom} g$, explicit

Weak and strong duality

Weak duality: $d^\star \leq p^\star$

- always holds (for convex and nonconvex problems)
- can be used to find nontrivial lower bounds for difficult problems

Strong duality: $d^\star = p^\star$

- does not hold in general
- holds for convex problems under **constraints qualifications**

Strong duality holds for a convex problem

$$\begin{aligned} \min_{x \in \mathcal{D}} \quad & f(x) \\ \text{subject to} \quad & f_i(x) \le 0, \quad i = 1, \dots, m \\ & Ax = b, \end{aligned}$$

if it is strictly feasible i.e.

$$\exists x \in \operatorname{int}\mathcal{D}: \quad f_i(x) < 0, \quad i = 1, \dots, m, \quad Ax = b$$

- also guarantees that the dual optimum is attained (if $p^\star > -\infty$)
- can be sharpened: e.g. can replace int$\mathcal{D}$ with relint$\mathcal{D}$ (interior of the affine hull); linear inequalities do not need to hold with strict inequality, ...
- other constraint qualifications exist e.g. LICQ

Example: inequality form LP

Primal problem

$$\min_{x\in\mathbb{R}^n} \quad c^{\mathrm{T}}x$$
$$\text{subject to} \quad Ax \le b$$

Dual function

$$g(\lambda) = \inf_{x\in\mathbb{R}^n} \left((c + A^{\mathrm{T}}\lambda)^{\mathrm{T}}x - b^{\mathrm{T}}\lambda\right) = \begin{cases} -b^{\mathrm{T}}\lambda, & A^{\mathrm{T}}\lambda + c = 0 \\ -\infty, & \text{otherwise} \end{cases}$$

Dual problem

$$\max \quad -b^{\mathrm{T}}\lambda$$
$$\text{subject to} \quad A^{\mathrm{T}}\lambda + c = 0, \quad \lambda \ge 0$$

- From Slater's condition: $p^\star = d^\star$ if $\exists\tilde{x}: \; A\tilde{x} < b$
- In fact, $p^\star = d^\star$ except when primal and dual are infeasible

Example: Quadratic program

Primal problem (assume P symmetric positive definite)

$$\min_{x \in \mathbb{R}^n} \quad x^{\mathrm{T}} P x$$
$$\text{subject to} \quad Ax \leq b$$

Dual function

$$g(\lambda) = \inf_{x \in \mathbb{R}^n} \left(x^{\mathrm{T}} P x + \lambda^{\mathrm{T}} (Ax - b) \right) = -\frac{1}{4} \lambda^{\mathrm{T}} A P^{-1} A^{\mathrm{T}} \lambda - b^{\mathrm{T}} \lambda$$

Dual problem

$$\max \quad -\frac{1}{4} \lambda^{\mathrm{T}} A P^{-1} A^{\mathrm{T}} \lambda - b^{\mathrm{T}} \lambda$$
$$\text{subject to} \quad \lambda \geq 0$$

- From Slater's condition: $p^\star = d^\star$ if $\exists \tilde{x} : \ A\tilde{x} < b$
- In fact, $p^\star = d^\star$ always

Example: nonconvex problem with strong duality

Primal problem

$$\begin{array}{ll} \min_{x \in \mathbb{R}^n} & x^{\mathrm{T}} A x + 2 b^{\mathrm{T}} x \\ \text{subject to} & x^{\mathrm{T}} x \leq 1 \end{array}$$

$A \not\succeq 0$ is not positive definite.

Dual function

$$g(\lambda) = \inf_{x \in \mathbb{R}^n} \left(x^{\mathrm{T}} (A + \lambda I) x + 2 b^{\mathrm{T}} x - \lambda \right)$$

- unbounded below if $A + \lambda I \not\succeq 0$ or if $A + \lambda I \succeq 0$ and $b \notin \mathcal{R}(A + \lambda I)$
- otherwise minimised by $x = -(A + \lambda I)^{\dagger} b$: $g(\lambda) = -b^{\mathrm{T}} (A + \lambda I)^{\dagger} b - \lambda$

Dual problem

$$\begin{aligned} \max \quad & -b^{\mathrm{T}}(A+\lambda I)^{\dagger} b - \lambda \\ \text{subject to} \quad & A + \lambda I \succeq 0 \\ & b \in \mathcal{R}(A+\lambda I) \end{aligned}$$

and equivalent semidefinite program:

$$\begin{aligned} \max \quad & -t - \lambda \\ \text{subject to} \quad & \begin{bmatrix} A+\lambda I & b \\ b^{\mathrm{T}} & t \end{bmatrix} \succeq 0 \end{aligned}$$

Strong duality although primal problem is not convex (not easy to show).

KKT conditions revisited

Karush-Kuhn-Tucker conditions are satisfied at $x^\star, \nu^\star, \lambda^\star$ i.e.

$$\nabla f(x^\star) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x^\star) + \sum_{i=1}^{p} \nu_i \nabla h_i(x^\star) = 0, \tag{2a}$$

$$h_i(x^\star) = 0, \quad i = 1, \dots, p \quad \text{[primary constraints]} \tag{2b}$$

$$f_i(x^\star) \le 0, \quad i = 1, \dots, m \quad \text{[primary constraints]} \tag{2c}$$

$$\lambda^\star \succeq 0, \quad i = 1, \dots, m \quad \text{[dual constraints]} \tag{2d}$$

$$\lambda_i^\star f_i(x^\star) = 0, \quad i = 1, \dots, m \quad \text{[complementary slackness]} \tag{2e}$$

Necessary condition: If strong duality holds and x, ν, λ are optimal, then they must satisfy KKT conditions.

For *any* problem for which strong duality holds, KKT are necessary conditions.

KKT conditions for convex problem

Sufficient condition: If $x^\star, \nu^\star, \lambda^\star$ satisfy KKT conditions and the problem is convex, then $x^\star, \nu^\star, \lambda^\star$ are primal and dual optimal:

- from complementary slackness: $f(x^\star) = f(x^\star) + \sum_{i=1}^m \underbrace{\lambda_i f_i(x^\star)}_{=0} + \sum_{i=1}^p \underbrace{\nu_i h_i(x^\star)}_{=0} = \mathcal{L}(x^\star, \lambda^\star, \nu^\star)$
- $g(\lambda^\star, \nu^\star) = \inf_x \mathcal{L}(x, \lambda^\star, \nu^\star)$ and from the 1st order necessary condition and convexity of f we have that the minimum is attained at $x^\star$, hence $g(\lambda^\star, \nu^\star) = \mathcal{L}(x^\star, \lambda^\star, \nu^\star)$

Thus it follows that $f(x^\star) = g(\lambda^\star, \nu^\star)$.

If Slater's condition is satisfied:
$x^\star$ is optimal if and only if there exists $\lambda^\star, \nu^\star$ that satisfy KKT conditions

- recall that Slater implies strong duality, and that the dual optimum is attained
- generalises optimality condition $\nabla f(x) = 0$ for unconstrained problem